SRC TR 87-187

Annotated Bibliography of Blocking Systems

by

L. GÜN

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **1987**	2. REPORT TYPE	3. DATES COVERED **00-00-1987 to 00-00-1987**

4. TITLE AND SUBTITLE **Annotated Bibliography of Blocking Systems**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Maryland,Electrical Engineering Department,College Park,MD,20742**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES **30**	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**			

ANNOTATED BIBLIOGRAPHY OF BLOCKING SYSTEMS

by

Levent GÜN*
Electrical Engineering Department
and
Systems Research Center
University of Maryland
College Park, MD., 20742

ABSTRACT

Queueing systems subject to blocking have been studied by researchers from different research communities. Due to the wide applicability of these models there is an exhaustive list of related papers. This annotated bibliography summarizes 55 of these papers by briefly describing the model, the performance measure(s), the basic methodology used and/or the results obtained in each paper.

* This research was supported by the Office of Naval Research through Grant N00014-84-K-0614 and through a grant from AT&T Bell Laboratories.

1. INTRODUCTION

Majority of the blocking systems considered in the literature are composed of a series arrangement of service stations (*nodes*) with *finite* capacity intermediate buffers between these stations. Because of the physical limitations on the buffer spaces and the variations in the service times, the flow of jobs through the system may get *blocked.* A typical *tandem* configuration is shown in Figure 1. At each node some work is performed on a job which is then passed on to the next node and finally ejected from the last station. Recently, the methods used in analyzing such tandem systems are extended to the analysis of open and closed queueing networks, denoted by OQN and CQN, respectively, hereafter.

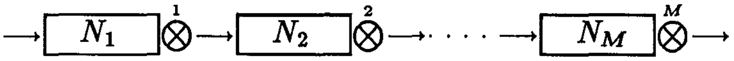

Figure 1

The literature considers two distinct blocking policies for transfer lines with finite buffers. Let S be one of the servers in Figure 1. Under the first policy, called *immediate* blocking hereafter, at a time of service completion, the server S is blocked if the downstream buffer becomes full due to this service completion; the server S remains blocked until the congestion is reduced downstream, at which time it resumes service and *begins* to process its next job (if any). Under the second policy, called *non-immediate* blocking hereafter, the server S is blocked at a service completion time if the job that has just completed service cannot proceed to the next buffer due to congestion. When the congestion is reduced downstream, this job proceeds to the next buffer without receiving any further service at server S, and the server S resumes service and begins processing its next job (if any). For general networks, it is possible that a buffer accepts inputs from several servers. When such a buffer is full, several upstream servers may be blocked at a given time. In such cases, the blocked servers are assumed to unblock on an *first-block-first-enter* basis, i. e., if, both servers i and j are blocked because of node k and server i is blocked before server j, when there is a departure from node k first the server i unblocks and server j remains blocked.

The papers reviewed here are classified into two major classes depending on whether the servers are reliable or subject to breakdowns.

(i) Systems with reliable servers : In this class, each node in the network is attended by *reliable* server. As long as there is a job to process and the server is not blocked, it can always give service according to some known service distribution. Some models allow an infinite capacity input buffer upstream of the first node server and assume jobs to arrive according to some statistical pattern. Others assume that there is an inexhaustible supply of jobs to the first node server, so that this server is *saturated* with jobs. With few exceptions, servers are assumed to have exponentially distributed service times

(ii) Systems with failure type servers : In this class, each node is attended by *unreliable* servers, i. e., servers subject to failures that are non-deterministic in both occurrence and duration. Some models assume *operational* failures, i. e., a server can only fail when working on a job. Others assume that the failures are *time dependent* only and independent from the state of the server. This class of models can be further divided into two subclasses according to the assumptions imposed on the service times.

(ii.a) Deterministic processing times : The service times are assumed to be *deterministic* while both the failure and repair times of the servers are allowed to be *random.* All the models discussed in this bibliography postulate that the first node server is saturated, and that the duration of up and down times are *geometrically* distributed.

(ii.b) Random processing times : Service times as well as the failure and repair times are *random* and, with a few exceptions, assumed to be exponentially distributed.

A common assumption to all the models surveyed is that the last node server(s) is never blocked, i. e., there is always space available into which the last stage server(s) can discharge a part. All models assumes a *single* job class and unless otherwise mentioned, it is assumed that a *single* server which operates according to the *first-come-first-served* queueing discipline is in attendance at each node. Operational failures are specifically mentioned in the model description while the failure type is not mentioned for time dependent failures. The papers in each class are presented in *historical order* as they appeared in the literature and an author index is provided at the end.

Finally, the survey given here is far from being complete and combines only the related papers that were available to the author at the time of the writting of this bibliography, any omissions are, off-course, unintentional. Indeed, the author will be more than happy to receive relevant papers for possible future updates of this annotated bibliography.

2. THE ANNOTATED BIBLIOGRAPHY

2.1. MODELS WITH RELIABLE SERVERS

[1] G.C. Hunt, " Sequential arrays of waiting lines", *Operations Res.*, **4**, pp. 674-683, 1956.

Model : A tandem line of *exponential* servers is considered with *Poisson* arrivals to an infinite capacity buffer in front of the first node server. The following models are discussed under the *non-immediate* blocking policy; (i) infinite capacity buffers between the servers, (ii) no buffers between the servers, (iii) finite capacity buffers between the servers, and (iv) a line of servers where the line moves at once as a unit (the unpaced belt production line) where no buffers and no vacant servers are allowed.

Measures : The mean number of jobs in the system and the maximum system utilization are considered.

Results : Maximum possible utilization is obtained in all four cases and graphs of the utilization vs. mean number of jobs in the system are displayed. It is concluded that for utilizations less than .5 blocking has very little effect on the system performance.

[2] H.S. Hillier and R.W. Boling, " Finite queues in series with exponential or Erlang service times - A numerical approach", *Operations Res.*, **15**, pp. 286-303, 1967.

Model : A tandem line of servers with *exponential* or *Erlang* service time distributions, separated by finite capacity buffers is considered under the *non-immediate* blocking policy. The first node server is assumed to be *saturated.*

Measures : The steady-state output rate and the mean number of jobs in the system are considered.

Method : When the service times have Erlang distribution the states are identified and the balance equations are solved numerically using the Gauss-Seidel method. For exponential service times an approximate procedure which analyzes each node individually as an $M/M/1/N$ queueing system is given.

[3] N.P. Rao, " Two-stage production systems with intermediate storage", *AIEE Trans.*, **7**, pp. 414-421, 1975.

Model : A two node tandem system with a finite capacity intermediate buffer is discussed. Service times are assumed to be *independent* and *exponentially* distributed for the first node server and have either *Erlang* or *normal* distribution for the second node server. The *non-immediate* blocking policy is assumed with a *saturated* first node server.

Measure : The effect of unbalancing on the mean production rate is considered.

Results : The equations for the steady-state queue size probabilities are shown to involve Laplace transforms of the density functions of the service time distributions and their derivatives. A recursive solution for the mean production rate is obtained and the effect of balancing is discussed. It is concluded that the balanced division of the total work among the servers is not optimum for an unbalanced system and slightly higher load is needed for the server with less variable service time distribution in order to maximize the production rate, and that the effect of unbalancing increases with the difference in the variabilities of the service time distributions.

[4] A.G Konheim and M. Reiser, " A queueing model with finite waiting room and blocking", *J. Assoc. Comput Mach.*, **23**, pp. 328-341, 1976.

Model : A two node tandem system with a finite capacity intermediate buffer and a *Poisson* arrival stream to an infinite capacity buffer in front of the first node server is considered. The output of the system is also fed back to the first buffer with a certain probability. Service times are assumed *independent* and *exponentially* distributed and that *immediate* blocking strategy is adopted.

Measures : The steady-state queue size probabilities are obtained.

Method : The state of the system at time t is defined by the pair (X_t^1, X_t^2), where X_t^1 and X_t^2 denotes the number of jobs in the first and the second buffer at time t, respectively. Forward equations are written and solved for the steady-state probabilities by using generating functions. Necessary and sufficient conditions are given for the stability of the system. Analytical results are then put into an algorithmic form and some special cases are discussed.

[5] A.B. Clarke, "A two-server queueing system with storage between servers", Math. Rep. **50**, Western Michigan University, 1977.

Model : A two node tandem system with *independent* and *exponential* service time distributions is considered with a finite capacity intermediate buffer between the servers. The *non-immediate* blocking policy is assumed with *Poisson* arrivals to an infinite capacity buffer in front of the first server. When both servers are idle, an incoming job goes directly to the second node, whereas if the second node server is busy but the first node server is idle, the job receives service in the first node server and then joins the intermediate buffer. When a service completion occurs in the second node, all the jobs in the intermediate buffer, including the one that is blocked in the first node (if any), leave the system. If at the time of a service completion in the first node, the second node server is idle, then the job that has just been served in the first node leaves the system without receiving any service in the second node.

Measures : Steady-state probabilities of the system state process are studied.

Method : A system state process is defined and the generator matrix is written explicitly. Then a matrix-geometric solution is given for the steady-state probabilities where the rate matrix is obtained as the minimal solution of a third order matrix equation. Some computational methods are also discussed.

[6] P. Caseau and G. Pujolle, " Throughput capacity of a sequence of transfer lines with blocking due to finite waiting room", *IEEE Trans. Software Engrg.*, **SE-5**, pp. 631-642, 1979.

Model : A tandem line of servers with *independent* and *exponential* service time distributions is considered. The servers are separated with finite capacity buffers. The *immediate* blocking policy is assumed with *Poisson* arrivals to an infinite capacity buffer in front of the first server. Two extensions are considered; (i) the service times depend on the queue sizes, and (ii) there are external *Poisson* arrivals of same parameter to the intermediate buffers. In the latter case, an intermediate arrival is rejected if the buffer is full, and if accepted, it leaves the system as soon as its service is completed.

Measure : Maximum system throughput is considered.

Method : The tandem line is replaced by several isolated $M/M/1/N$ queueing systems with equivalent arrival and service rates by using the equivalence relations between different types of blocking policies. A numerical algorithm that uses recursive relations for the utilizations of these subsystems is proposed. Exact expressions for the maximum throughput is obtained when (i) there are only two nodes, (ii) the intermediate buffer capacities are equal and the servers are identical.

[7] F.G. Foster and H.G. Perros, " On the blocking process in queueing networks", *European J. Operations Res.*, **5**, pp. 276-283, 1980.

Models : Three different models, all with *exponential* servers are discussed under the *non-immediate* blocking strategy. The first model is a two node tandem system with a finite capacity intermediate buffer and a *Poisson* arrival stream to an infinite capacity buffer in front of the first node server. The second model is again a two node tandem system but with several servers in parallel in the first node and a single server in the second node with no buffer space between them. There are *Poisson* arrivals to infinite capacity buffers in front of each one of the first node servers. In the third model, two systems of the type described for the second model are in parallel in the first node, which is in tandem with a single server, again with no intermediate buffer.

Measure : The mean blocking time is considered.

Results : Exact expressions for the mean blocking time is obtained for the first two models when the first node servers have (i) infinite rates, and (ii) minimum rates (to insure stability). For the third model approximate results are given.

[8] J. Labetoulle and G. Pujolle, " Isolation method in a network of queues", *IEEE Trans. Soft. Engrg.*, **SE-6**, pp. 373-381, 1980.

Model : An OQN with *exponential* servers and finite capacity intermediate buffers is considered with a *Poisson* arrival process to the network under the *immediate* blocking policy.

Method : An isolation method is presented where the network is subdivided into several subsystems so that each system can be studied independently. The method is said to give powerful approximations in cases where the service times or the arrival rates depend on the state of some part of the queueing network.

[9] G. Latouche and M.F. Neuts, " Efficient algorithmic solutions to exponential tandem queues with blocking", *SIAM J. Alg. Disc. Meth.*, **1**, pp. 93-106, 1980.

Model : A two node tandem system with a finite capacity intermediate buffer is considered. There are r and c *identical* parallel servers in the first and the second node, respectively. All the servers are assumed to have *independent* and *exponentially* distributed service times. There is a *Poisson* arrival stream to an infinite capacity buffer in front of the first node. The *non-immediate* blocking strategy is adopted, with *full* blocking of all the first node servers in that when r^* servers in the first node are blocked, all the rest are also blocked. First node servers resume service when the number of departures from the second node, after the full blocking of the first node, reaches k^*.

Measures : Steady-state probabilities for the joint queue length distribution are considered.

Method : A system state process is defined and explicit expressions for the block entries of the generator matrix are given for different choices of r^* and k^*. The stationary probability distribution vector is obtained in matrix-geometric form. Some extensions and variants of the above model are briefly mentioned.

[10] Y. Takahashi, H. Miyahara and T. Hasegawa, " An approximation method for open restricted queueing networks", *Operations Res.*, **28**, pp. 594-602, 1980.

Model : An OQN with finite capacity buffers is considered when the service times are *exponentially* distributed while the arrival process is *Poisson.* The *non-immediate* blocking policy is assumed.

Measures : Blocking probabilities and the output rates are considered.

Method : An approximate decomposition method is proposed through the introduction of a pseudo-arrival rate and an effective service rate, so that each node can be treated as an $M/M/1/N$ system in isolation.

[11] H.G. Perros, " A symmetrical exponential open queue network with blocking and feedback", *IEEE Trans. Software Engrg.*, **SE-7**, pp. 395-402, 1981.

Model : A two node tandem system with feedback is considered. There are several homogeneous servers in parallel in the first node while only a single server in the second node, all with *exponential* service time distributions. *Identical* and *independent Poisson* processes feeds infinite capacity buffers in front of the first node servers. A first node server gets blocked at *each* service completion time and remains blocked until the job that it last served completes its service in the second node server.

Measure : Queue-length distribution of the first node buffers is studied.

Method : An approximate expression for the probability distribution of the number of blocked first node servers is first obtained. Based on this distribution and assuming processor sharing type of service, an approximate expression is derived for the queue length probability distribution.

[12] M. Pinedo and R.W. Wolff, " A comparison between tandem queues with dependent and independent service times", *Operations Res.*, **30**, pp. 464-479, 1982.

Model : A two node tandem system with *Poisson* arrivals to an infinite capacity buffer is considered under light traffic conditions when there is infinite or no buffer space between the servers. The service times are either *independent* and *exponentially* distributed at each server (case I), or are generated according to an exponential distribution and are *same* at each server (case D). Two node tandem configurations with general service time distributions with a *saturated* first node server is also considered when the intermediate buffer capacity is either infinite or zero.

Measures : The expected waiting time $E(W)$, the mean and the variance of the k^{th} departure epoch $E(U_k)$ and $V(U_k)$, respectively, and the system capacity λ_{sup}, are compared for cases (D) and (I). The effect of the service time regularity on the performance of the system is also considered. The notations $X(I)$ and $X(D)$ are used to denote the performance measure X under cases I and D, respectively.

Results : Both $E(W(D))$ and $E(W(I))$ are computed for the infinite and zero capacity intermediate buffer between the servers, both under light traffic. It is concluded that under light traffic $E(W(D)) > E(W(I)))$ when the intermediate buffer has infinite capacity. As for the departure epochs, it is shown that $E(U_k(D)) \leq E(U_k(I))$ for an infinite capacity intermediate buffer, while $E(U_k(D)) = E(U_k(I))$ and $V(U_k(D)) \geq V(U_k(I))$ for $k \geq 1$, when there is no intermediate buffer. Expressions for λ_{sup} are obtained for a general tandem line where servers with *general* service time distributions are in attendance. For the two server tandem system it is shown that $\lambda_{sup}(D) = \lambda_{sup}(I)$, while for a general system $\lambda_{sup}(D) < \lambda_{sup}(I)$, provided that the service times are not deterministic. Effect of the service time regularity on the performance of the tandem system is also considered. It is shown that a tandem system that involves servers with less variable service time distributions has a larger capacity compared to ones with more variable service time distributions. Based on both analytical and simulation results it is concluded that the *relative* performance of case (I) improves as (i) the arrival rate decreases, (ii) the arrival process becomes more regular, and (iii) the service time distribution becomes less regular.

[13] T. Altıok, " Approximate analysis of exponential tandem queues with blocking", *European J. Operations Res.*, **11** , pp. 390-398, 1982.

Model : A tandem line of servers with *independent* and *exponential* service time distributions is considered with finite capacity intermediate buffers. The *non-immediate* blocking policy is adopted with *Poisson* arrivals to an infinite capacity buffer in front of the first server.

Measure : An approximate algorithm for calculating the steady-state queue length probabilities is given.

Method : A decomposition procedure that revises the service time distribution of each server and decomposes the system into isolated simple queueing systems is presented. Decomposition is done by ignoring the interactions between the components of the system. Two basic assumptions are made; (i) the input process to each intermediate buffer is assumed to be Poisson, and (ii) the blocking of a server occurs only due to the immediate successor buffer. The method decomposes the tandem line into queueing systems of the type $M/C_2/1/N$, where C_2 denotes a two stage Coxian distribution. A Markov chain is imbedded by looking at each system at departure points. Analytic results lead to systems of equations which are solved by iteration techniques. Numerical results and some extensions are mentioned.

[14] F.P. Kelly, " The throughput of a series of buffers", *Advances in Appl. Probability*, **14**, pp. 633-653, 1982.

Model : Messages has to be transmitted through a tandem channel with M nodes. The nodes have finite but *equal* buffer capacities. Lengths of the messages are *independent and identically distributed (i.i.d.)* with a known common distribution. Time taken by a node to transmit a message (service time) is proportional to the message length and a given message has *same* transmission time at each node. Inputs to the first node are assumed instantaneous, i. e., *saturated* first node server. Both the *immediate* and *non-immediate* blocking policies are discussed. The model where the transmission rates of a message at different nodes are not equal but independently distributed according to some distribution is also discussed.

Measure : Asymptotic behavior of the system throughput is studied as the number of nodes increases.

Results : The throughput is defined as $\lim_{t\to\infty} E(\frac{N_t}{t})$, where N_t is the number of messages which have been transmitted from the first node in the interval $[0,t]$. First, the growth rate of the intermediate buffer sizes in order the throughput not to decrease to 0 as $M \to \infty$ is investigated. Systems with no intermediate buffers are used to provide straightforward bounds on the degradation of the throughput as the number of nodes increase. More complex bounds are obtained in order to study the effect of an increase in the buffer capacities. It is shown that for exponentially distributed message lengths, either the transmission rate or the buffer capacity should grow at rate *log* M, while for distributions with tail parts proportional to $x^{-\rho}$, $\rho > 1$, either the transmission rate should grow at a rate $M^{\frac{1}{\rho}}$ or the buffer capacities should grow at a rate $M^{\frac{1}{(\rho-1)}}$, in order the throughput not to decrease to zero as $M \to \infty$.

[15] H.G. Perros and T. Altıok, " Approximate analysis of open networks of queues with blocking: Tandem configurations", CS Rep. **83-11**, NC State University, 1984.

Model : A tandem line of M servers with *independent* and *exponential* service time distributions and finite capacity intermediate buffers is considered. The *non-immediate* blocking policy is assumed with *Poisson* arrivals to an infinite or finite capacity buffer in front of the first node server.

Measure : An approximate algorithm for calculating the steady-state queue length probabilities is given.

Method : When the first buffer has infinite capacity, the decomposition method of [13] is used by relaxing assumption (ii) in that blocking is allowed to backlog over any number of successive queues. The system is now decomposed into several $M/C/1/N$ queueing systems in isolation, where C is a $M-i+1$ stage Coxian distribution for the i^{th} node, for $1 \leq i < M$. In the case where the first buffer has finite capacity, the effective arrival rate is estimated by successive iterations. Numerical examples are given and the approximation is reported to be better for balanced systems.

[16] H.G. Perros, " Queueing networks with blocking: A bibliography", *Sigmetrics Newsletter*, pp. 8-11, Spring 1984.

An exhaustive list of papers on the analytical and numerical investigations of the queueing networks with blocking is compiled.

[17] F.P. Kelly, " Segregating the input to a series of buffers", *Math Operations Res.,* **10**, pp. 33-43, 1985.

Model : Two parallel systems of the type described in [14] is considered where the incoming message is directed into one of the two systems depending on the message length.

Measure : System throughput is studied.

Result : It is shown that by using two systems in parallel, one dealing with long messages and the other with short messages, the decay of throughput with the number of nodes M can be improved from $(logM)^{-1}$ to $(loglogM)^{-1}$, when the message length distribution is exponential, while for message length distribution with tail part proportional to $x^{-\rho}$, with $\rho > 1$, the improvement is from $M^{-1/\rho}$ to M^{-1/ρ^2}.

[18] Ç. Büyükkoç, " An approximation method for feed-forward queueing networks with finite buffers: A manufacturing perspective", manuscript, 1985.

Model : Various queueing models with finite capacity buffers and *exponential* service time distributions are considered under the *non-immediate* blocking policy. The arrivals to the system are assumed *Poisson.* The tandem configuration possibly with multiple servers, split and merge configurations and the triangle configuration are considered.

Measures : Average queue lengths at each node and the stability condition for the network are studied.

Method : An approximate decomposition procedure based on the flow conservation is proposed. The system is decomposed into several $M/M/1/N$ queues, where the effective arrival rates are obtained by the flow conservation principle while the effective service rates are obtained as the solutions of fixed point problems.

[19] L. Kerbache and J.M. Smith, "The generalized expansion method for open queueing networks", manuscript, 1986.

Model : OCNs with finite capacity intermediate buffers are considered with *independent renewal* arrival processes and *general* service time distributions.

Method : An approximate decomposition approach based on the "Generalized Expansion" method is presented by focusing on the two stage Erlang or hyperexponential service time distributions. First, the network is expanded by adding a holding node to each finite buffer to register blocked jobs with appropriate routing probabilities. Then, in order to determine the parameters of the network, the first two moments of the service and the interarrival distributions are approximated throughout the expanded network. Finally, feedback routing arcs are eliminated to avoid strong arrival dependencies. Several numerical examples are given.

[20] H.G. Perros and P.M. Snyder, " A computationally efficient approximation algorithm for analyzing open queueing networks with blocking", *Technical Report,* **TR-86-13**, North Carolina State University, Raleigh, N.C., 1986.

Model : An OQN with finite or infinite capacity buffers, *Poisson* external arrivals and *exponential* service time distributions is considered. Finite buffers are not allowed to accept both inputs from other servers and external arrivals. The network is assumed deadlock free and the *non-immediate* blocking policy is adopted under the *first-blocked-first-enter* rule.

Measure : The steady-state queue length distibution is considered.

Method : An approximation algorithm is presented where the network is decomposed into individual queues by revising the arrival and service processes and the capacities of the finite buffers. A node with a buffer of capacity N is approximated by a $M/C_2/1/N+K$ queueing system in isolation, where K is the the number of upstream queues linked to this buffer and C_2 is a two stage Coxian distribution and incorporates the effect of blocking. The parameter of the Poisson arrival distribution to each queue is obtained by the flow conservation principle. For the case where the first buffer has finite capacity, the effective arrival rate is estimated by successive iterations. Numerical examples are given.

[21] R.O. Onvural and H.G. Perros, "Some exact results on closed queueing networks with blocking", *Technical Report,* **TR-86-14,** North Carolina State University, Raleigh, N.C., 1986.

Model : A CQN with finite capacity buffers and *exponential* service time distributions is considered. Deadlocks are assumed to be detected and resolved immediately. The *non-immediate* blocking policy is adopted under the *first-blocked-first-enter* rule.

Measures : The steady-state queue length distibution and the system throughput are considered.

Method : First, results that relate the throughput of a CQN with finite capacity buffers to the throughput of a CQN with infinite capacity buffers are given and equivalences between CQNs with finite capacity buffers with respect to the buffer capacities are obtained. Then, exact numerical algorithms for analyzing three special CQNs with symmetric queues are provided.

[22] H.G. Perros, A.A. Nilsson and Y.C. Liu, "Approximate analysis of product-form type queueing networks with blocking and deadlock", *Technical Report,* **TR-86-19**, North Carolina State University, Raleigh, N.C., 1986.

Model : A product-form type CQN in which some of the buffers have finite capacities is considered. The service time distributions are *exponential* and deadlocks are assumed to be detected and resolved immediately. The *non-immediate* blocking policy is adopted under the *first-blocked-first-enter* assumption.

Measures : The steady-state queue length distribution and the system throughput are considered.

Method : First, when all the buffers have finite capacity, an algorithm that generates all the states and the rate matrix Q of the system is given. The algorithm uses the Gauss-Seidel iteration method to solve the equation $xQ = 0$. This algorithm is then used in the approximate analysis of CQNs when some of the buffers have finite capacities. The analysis is based on the Norton's theorem but is not efficient when the blocking subnetwork, i. e., the part of the network that includes the finite queues and the infinite queues that are liable to blocking, is large.

[23] L. Gün and A.M. Makowski, " An approximation method for general tandem queueing systems subject to blocking", submitted to the *Workshop on Queueing Networks with Blocking,* NC State University, 1988, original manuscript 1986.

Model : A tandem line of servers with finite capacity intermediate buffers and *phase-type* service time distributions is considered. Both the *immediate* and the *non-immediate* blocking policies are discussed with a *saturated* first node server.

Measure : The steady-state marginal queue length distribution of each buffer is considered.

Method : An iterative algorithm where the tandem line is decomposed into several two node subsystems is presented. The effective service time distributions of the first and the second node servers are shown to be again phase-type with a phase structure that includes all the phases of the upstream and the downstream servers, respectively, to capture the effect of idling and blocking, respectively. These two node subsystems are solved efficiently by using the exact results obtained in (23). The algorithm is suitable for parallel computation. Numerical examples are provided.

[24] L. Gün and A.M. Makowski, "Matrix-geometric solution for finite capacity queues with phase-type distributions", to appear in the *Proceedings of Performance'87,* Brussels, Belgium, 1987.

Model : A class of finite state Quasi-Birth-and-Death (QBD) processes, that contains several well-known queueing models is considered. Two node closed queueing systems and two node tandem systems with a *saturated* first node server, both with finite buffers, feedbacks and *phase-type* service time distributions are among the models discussed in this class.

Measures : The invariant probabilities of the joint queue length are studied.

Method : An *explicit* solution is presented in *matrix-geometric* form for the invariant probability vector of the underlying Markov chain. No assumptions are made on the irreducibility of the QBD process. An easily computable expression is given for a rate matrix and several computational issues are briefly discussed.

[25] L. Gün, " Closed-form matrix geometric solution for a class of Quasi-Birth-and-Death processes", *Proceedings of the* 21st *Annual Conference on Information Sciences and Systems,* John Hopkins University, Baltimore, MD, 1987.

Model : Another class of finite state QBD processes, that contains several other well-known queueing models is considered. Two node tandem systems with a finite capacity intermediate buffer, feedbacks and *Poisson* arrivals to a finite capacity buffer in front of the first node with (i) *phase-type* servers and (ii) homogeneous *parallel exponential* servers at each node are among the models included in this class.

Measures : The invariant probabilities of the joint queue length are studied.

Method : Under no assumptions on the irreducibility of the QBD process, the invariant probability vector is shown to admit the *matrix-geometric* property. The method is based on a simple observation and provides an explicit expression for a rate matrix.

[26] R.O. Onvural and H.G. Perros, "Some exact results on closed queueing networks with blocking", manuscript, 1987.

Model : Cyclic CQNs with finite capacity buffers and *exponential* service time distributions are considered under both the *immediate* and the *non-immediate* blocking policies.

Measure : The system throughput is studied.

Method : Based on the numerical evidence, the throughput curve, as a function of the number of jobs K in the system, is observed to increase up to a value K^* and then decrease monotonically, and that the curve is symmetric for systems operating under the immediate blocking policy but unsymmetric for the non-immediate blocking policy. For the case of immediate blocking, K^* is obtained easily by using this symmetry. For the non-immediate blocking case, K^* is obtained by using some equivalence relations between the two blocking types for cyclic networks. These results are then used to approximate the throughput curve by fitting a curve (resp. two curves) through several known points in the immediate (resp. non-immediate) blocking case.

[27] K.P. Jun and H.G. Perros, "An approximate analysis of open tandem queueing networks with blocking and general service times", manuscript, 1987.

Model : The model of [15] is considered with two stage Coxian service time distributions.

Measures : The steady-state queue length probabilities are considered.

Method : An algorithm where the tandem system is decomposed into individual queues in isolation by revising the service and the arrival processes and the buffer capacities is presented. The revised service and the arrival processes are chosen to have two stage Coxian distributions whose parameters are computed using an iterative scheme so as to approximate the effect of the blocking and idling, respectively. Approximating the arrival process by a two stage Coxian distribution instead of a Poisson process is reported to give significantly better results.

[28] I.F. Akyıldız,“ On the exact and approximate throughput analysis of closed queueing networks with blocking", to appear in *IEEE Transactions on Software Engineering*, 1987.

Model : An M node CQN with *exponential* servers is considered with finite or infinite capacity buffers. The network is assumed *deadlock* free and the *non-immediate* blocking policy is assumed with the *first-blocked-first-enter* rule.

Measures : The steady-state queue length distribution and the throughput of each node are considered.

Method : An *exact* product-form solution is obtained for two node CQNs by finding an equivalent nonblocking network with the same state space. Such a state space transformation is not possible for $M > 2$, and an *approximate* transformation is given in this case so that the number of states in both networks are approximately equal. This guarantees that the Markov processes describing the evaluation of both networks have approximately the same stochastic structure. Several numerical examples are given.

In the next two papers, the methodology of [28] is extended to CQNs where multiple exponential servers with the same parameter are in attendance at each node.

[29] I.F. Akyıldız, "Exact product form solution for queueing networks with blocking", *IEEE Transactions on Computers,* **C-36**, pp. 122-125, 1987.

[30] I.F. Akyıldız, "Product form approximations for closed queueing networks with multiple servers and blocking", to appear in *IEEE Transactions on Software Engineering*, 1987.

[31] I.F. Akyıldız, "Mean value analysis for blocking queueing networks", to appear in *IEEE Transactions on Software Engineering*, 1987.

Model : The CQN described in [28] is considered.

Measures : The mean number of jobs, the mean waiting time, the throughput and the blocking probability of the i^{th} node are considered.

Method : An iterative approximation algorithm that utilizes the mean value analysis for queueing networks with blocking is proposed. The approximations are based on the modification of the mean residence times due to the blocking events that occur in the network.

[32] I.F. Akyıldız, "Analysis of reversible and nonreversible queueing networks with rejection blocking", manuscript, 1987.

Model : The CQN described in [28] is considered under the *rejection* blocking policy, in that the job blocked under the non-immediate blocking policy immediately starts receiving another service, and this is repeated until the server of the downstream node releases a job.

Measures : The mean number of jobs and the throughput of each node are considered.

Method : First, for CQNs with reversible routing, the known exact product form solution for the steady-state probabilities is presented and an algorithm for the computation of the normalization constant is given so that the computation of the above performance measures is possible. Then, for reversible networks with buffer capacity N_i at node i, the blocking network is shown to be isomorphic to a nonblocking network, by a state transformation, under the condition $N_i > \sum_{i=1}^{M} N_i - K$, $1 \leq i \leq M$, where M and K are the number of nodes and the number of jobs in the network, respectively. Based on this equivalence the product form solution is possible and the exact formulas are given for computing the performance measures.

2.2. MODELS WITH UNRELIABLE SERVERS

(1) M.C. Freeman, " The effects of breakdowns and internode storage on production line capacity", *J. Industrial Engrg.*, **151**, pp. 194-200, 1964.

Model : A tandem line of servers with *constant* and *equal* services times is considered. The *non-immediate* blocking policy is assumed with a *saturated* first node server. Failures are assumed operational in that a server can only fail when working on a job. The times between successive breakdowns and the duration of breakdowns, i. e., the *up* and *down* times, respectively, are all assumed to be *independent* and *exponentially* distributed.

Measure : The line efficiency, defined as $\frac{P_D(0)-P_D(N)}{P_D(0)-P_D(\infty)}$, is studied, where $P_D(N)$ is the percentage of the time the line is down for a given buffer capacity N.

Results : First, $P_D(0)$ and $P_D(\infty)$ are calculated. Then, the effect of system parameters on the performance of the system is discussed through simulation results for three node systems. Some general guidelines based on the simulation results are given for capacity allocation of buffers.

(2) J. Masso and M.L. Smith " Internode storages for three node lines subject to stochastic failures", *AIIE Trans.*, **6**, pp. 354-358, 1974.

Model : A three node tandem system with finite capacity intermediate buffers is considered when the service times are *equal* and *constant*, while the up and down times are *independent* and *exponentially* distributed. The first node server is assumed *saturated.*

Measure : System utilization is considered as the performance measure.

Method : Single and multiple regression techniques are used to approximate the minimal total buffer capacity required in order the system to reach its maximal possible utilization level. A technique to allocate a given quantity of total capacity among individual internode buffers is also given.

Results : Simulations showed that the effect of increasing the buffer capacity on the system utility is insignificant when the system utility is within 5% of its maximal value.

(3) T.J. Sheskin "Allocation of internode storage along an automatic production line," *AIIE Trans.,* **8**, pp. 146-152, 1976.

Model : A tandem line of servers with *constant* and *equal* production times is considered. The *immediate* blocking policy is assumed with the first node server *saturated.* The up and down times are assumed to be *independent* and *exponentially* distributed. Failures, repairs and transfer of jobs are all synchronized to certain time epochs.

Measure : Allocation of a fix total buffer capacity is considered so as to maximize the steady-state output rate.

Method : An *exact compression* algorithm is given for two node tandem systems. For larger systems, a much faster *approximate decomposition* algorithm is proposed. This algorithm analyzes each server separately by ignoring the dependence between the arrivals and departures to a node.

(4) G.T. Artamanov, " Productivity of a two instrument discrete processing line in the presence of failures", *Kibernatika* **3**, pp. 126-130; English trans. *Cybernetics*, **12** , pp. 464-468 1977.

Model : Two servers in tandem with *equal* and *constant* service times and an intermediate finite capacity buffer is considered. The *immediate* blocking policy is adopted with the first node server *saturated.* The up and down times are assumed to be *independent* and *exponentially* distributed.

Measure : The mean productivity is calculated using the steady-state probabilities.

Method : A continuous-time Markov chain whose states are defined by the triplet (n, α_1, α_2) is considered, where n is the number of jobs in the intermediate buffer and α_1, α_2 are the up/down indicators for the first and the second server, respectively. The balance equations for the steady-state probabilities are written explicitly and closed form solutions are obtained.

(5) E. Ignall and A. Silver, " The output of a two-node system with unreliable machines and limited storage", *AIIE Trans.*, **9**, pp. 183-188, 1977.

Model : A two node tandem system with a finite capacity intermediate buffer is considered with multiple servers at each node. The servers are assumed to have *constant* and *equal* service times and *independent* and *exponentially* distributed up and down times with *operational* failures. The *non-immediate* blocking policy is assumed with the first node servers *saturated.*

Measure : A computationally simple heuristic procedure for estimating the hourly line output is given.

Method : The line output is obtained approximately when there is a single server at each node, by using the known results for zero and infinite capacity buffers. For the case of multiple servers, each node is modeled as having a single server whose rate is equal to the sum of the individual rates of the servers.

(6) K. Okamura and H. Yamashina, " Analysis of the effect of buffer storage capacity in transfer line systems", *AIEE Trans.*, **9**, pp. 127-135, 1977.

Model : A two node tandem system with a finite capacity intermediate buffer is considered. The servers are assumed to have *constant* and *equal* service times, and *independent* and *geometrically* distributed up and down times with *operational* failures. The *non-immediate* blocking policy is assumed with a *saturated* first node server.

Measures : The effect of the buffer capacity on the production rate and the mean number of jobs in the buffer are considered.

Results : The system states and the corresponding probability transition matrix are explicitly written and the corresponding balance equations are given for a numerical solution of the steady-state probabilities. Graphs for the production rate and the mean number of jobs in the buffer are illustrated as a function of the intermediate buffer capacity, and a classification of these graphs are made. The effect of the variations in the production times for unbalanced systems is also considered. It is argued that the difference between the breakdown rates reduces the effect of installing an extra buffer, while the difference between the repair rates does not, and although the effect of interchanging the servers is negligible, for large differences it is better to put the faster server in front.

(7) J.A. Buzacott and L.E. Hanifin, " Models of automatic transfer lines with inventory banks- A review and comparison", *AIIE Trans.*, **10** , pp. 197-207, 1978.

Models : The assumptions, method of derivation and the results of papers by Vladziewski and Sevastyanov, Koenigsberg, Buzacott and Sheskin are compared for two node blocking systems with *independent* and *exponentially* distributed service, up and down times. The major difference between the models is whether the idling machines can or cannot fail.

Measure : The line efficiency is compared for these models.

Results : In this mostly qualitative paper, validity of the assumptions are tested by a real data from a transfer line and the predictions of the analytical models are compared with a simulation model which uses the actual data. The difference between the analytical and simulation results is reported to be significant.

(8) R.A. Murphy, " Estimating the output of a series production system", *AIEE Trans.*, **10**, pp. 139-148, 1978.

Model : A tandem line with finite capacity intermediate buffers between *some* of the servers is considered. Up and down times are assumed mutually *independent* with *exponential* and *general* distributions, respectively. The servers are assumed to have *constant* but *different* service times. The failures are assumed *operational* and no simultaneous repairs are permitted. The *non-immediate* blocking policy is assumed with a *saturated* first node server.

Measure : The average output rate is considered.

Method : First, results are obtained for a tandem line with no intermediate buffers. Then, the case when there is only one intermediate buffer in the line is discussed by considering the servers in front of this buffer as an input block and the ones after the buffer as an output block. Effective up and down times are first calculated and then, approximations are made to simplify the calculations so that sensitivity analysis and numerical optimization can be performed. The results are applied to the case where there are more than one intermediate buffers in the line, by finding the equivalent up and down times of the downstream servers, whence eliminating these buffers.

(9) Y.C. Ho, M.A. Eyler and T.T. Chien, " A gradient technique for general buffer storage design in a production line", *Proc.* 17th *IEEE Conf. on Control and Decision*, pp. 625-632, 1978.

Model : A tandem line of servers with *constant* but *different* service times is considered. The *non-immediate* blocking policy is assumed with a *saturated* first node server. The up and down times are assumed to be *independent* and *exponentially* distributed.

Measure : Allocation of a given total buffer capacity between the intermediate buffers in order to maximize the line efficiency is considered.

Method : An algorithm that generates the gradients of all buffers in a single simulation run is presented and simulation results are displayed. The algorithm first computes the sensitivity (gradient) of an increase of the line production per a unit increase in the buffer capacity at a buffer location, and then allocates a buffer size to each location by a hill climbing procedure until all gradients become equal.

(10) S.B. Gershwin and M. Ammar, " Reliability in flexible manufacturing systems", *Proc.* 18th *IEEE Conf. on Control and Decision*, pp. 540-545, 1979.

Model : Tandem and *merge* configurations are considered for systems with three servers. Both *deterministic* and *exponential* service time distributions are considered. Up and down times are assumed either to be *geometric* or *exponential* and *independent* from the state of the system. Failures are assumed *operational* and the *non-immediate* blocking policy is considered with a *saturated* first node server.

Measure : The steady-state queue length distribution is considered.

Method : For tandem systems a state process is defined and the state transition equations are written for the internal states both for exponential and deterministic (and equal) service times. Transition equations for a merge configuration are also written for the internal states and they are shown to be similar to the equations for the tandem model. Comparisons and some speculations are made for more complex merge configurations. In all cases, the steady-state probabilities for the internal states are assumed to be in sum-of-products form.

(11) J. Wijngaard, " The effect of internode buffer storage on the output of two unreliable production units in series, with different production rates", *AIIE Trans.*, **11**, pp. 42-47, 1979.

Model : A two node tandem system with a finite capacity intermediate buffer is considered. The servers are assumed to have *constant* but *different* service times and *independent* and *exponentially* distributed up and down time distributions. The *immediate* blocking policy is assumed with a *saturated* first node server. The main difference from the other models is that the "number" of jobs in the buffer is modeled as a *continuous* random variable rather than being discrete.

Measure : The system production rate is considered. The production rate is defined as quotient of the expected production per cycle and the expected duration of that cycle, where a cycle is the time between subsequent enterance points to states that corresponds to empty buffer.

Method : For equal production rates, the problem gave rise to a differential equation which is solved explicitly. For different production rates, a system of three differential equations is obtained and only the form of the solution is given. Simulation results for the effect of the intermediate buffer on an unbalanced line are also briefly discussed.

(12) A.L. Soyster, J.W. Schmidt and M.W. Rohrer, " Allocation of buffer capacities for a class of fixed cycle production lines", *AIEE Trans.*, **11**, pp. 140-146, 1979.

Model : A tandem line with finite capacity intermediate buffers between the servers is considered. The servers are assumed to have *constant* but *different* service times, while the down time distribution of each server is assumed to be an *independent Bernoulli* process. The *non-immediate* blocking policy is assumed with a *saturated* first node server.

Measure : Allocation of a given set of buffer capacities to maximize the steady-state output rate is considered.

Method : The problem is formulated as a nonlinear programming problem where the form of the objective function is unknown. The objective function is approximated by using lower and upper bounds. First, an exact expression is obtained for the two node case. For the general case, approximate expressions are given by using two server subsystems. Upper and lower bounds are established for the steady-state system output, and certain concave, separable programs are formulated in order to determine the optimal buffer capacities. Simulations showed that the objective function obtained through approximations is insensitive to modest changes in capacity allocation. It is also concluded that larger buffer capacities should be allocated around the less reliable servers.

(13) S.B. Gershwin and O. Berman, " Analysis of transfer lines consisting of two unreliable machines with random processing times and finite storage buffers", *AIIE Trans.*, **13**, pp. 2-11, 1981.

Model : A two node tandem system with *exponential* service times and a finite capacity intermediate buffer is considered. The servers are subject to *operational* breakdowns with *independent* and *exponentially* distributed up and down times. The *immediate* blocking policy is assumed with a *saturated* first node server.

Measures : The system production rate, utilization and the average queue length are considered.

Method : A continuous-time Markov chain is considered with state space as given in (4). Explicit expressions for the steady-state probabilities of these states are given by using the balance equations and assuming product-form solutions. These probabilities are then used to calculate the above performance measures, which are displayed as a function of the productivity of each server. The limiting behaviors are also discussed.

Results : The existence of a *saturation effect* is illustrated through numerical examples. The system saturates after a point in the sense that no further increase in the speed of the first node server can improve the productivity of the system. It is illustrated that the system's production rate increases as the productivity of the servers increase. The average queue length increases (resp. decreases) as the first (resp. second) node server becomes more productive.

(14) E.J. Muth and S. Yeralan, " Effect of buffer size on productivity of work stations that are subject to breakdowns", *Proc.* 20th *IEEE Conf. on Decision and Control*, pp. 643-648, 1981.

Model : A two node tandem system with a finite capacity intermediate buffer is considered. The service times are assumed to be *constant* and *equal* while the up and down times are *independent* and *exponentially* distributed. Only *operational* failures are allowed and the *non-immediate* blocking policy is assumed with a *saturated* first node server.

Measure : Variation of the system productivity with the intermediate buffer capacity is considered.

Method : A system state process is defined and $4N + 8$ states are identified, where N is the intermediate buffer capacity. The states are so ordered that the resulting transition matrix has a block tridiagonal structure with 4×4 blocks. This structured form leads to the solution of the steady-state probabilities by successively solving a system of 4 simultaneous equations. Moreover, the probabilities for the internal states are shown to have a *scalar-geometric* property. Specifically, if X is the random variable that denotes the number of jobs in the buffer at steady-state, then $P(X = k) = \lambda^k P(X = 1)$, for $1 < k < N$, where the scalar λ is an eigenvalue of a 4×4 matrix. The system production rate is calculated by using these probabilities and in general depends on several system parameters. In order to study the behavior of the system production rate as a function of the intermediate buffer capacity, a simpler, approximate expression that only depends on the intermediate buffer capacity is given by an empirical formula. This empirical formula is reported to be correct up to 10^{-10} over a wide range of parameters.

(15) T. Ohmi, " An approximation for the production efficiency of automated transfer lines with in-process storage", *AIEE Trans.*, **13**, pp. 22-28, 1981.

Model : A tandem line of servers with *constant* and *equal* service times is considered where groups of servers are mechanically interlocked with finite capacity buffers between each group. The first server is assumed *saturated.* Each server is assumed to have a *constant* probability of breaking down. Only *operational* failures are allowed and all servers in a group stops when one of its servers breakdowns. Repair times of each server are *i.i.d.* with common *exponential* distribution.

Measures : Production efficiency and the capacity allocation are considered.

Method : First, lifetime of each server is shown to have a *geometric* distribution. Then, a method for approximating the line efficiency is developed under the following assumptions; (i) only one group can be down at a given time, and (ii) at the time when a group breaks down the number of jobs in each buffer equals its line-averaged value, i.e., fluctuations of in-process inventories are ignored. The optimal partitioning of the line and a method for allocating capacities to the buffers are also numerically investigated. Hueristics based on the computational experiences are mentioned.

(16) T. Altıok and S. Stidham, Jr., " A note on transfer lines with unreliable machines, random processing times and finite buffers", *IIE Trans.*, **14**, pp. 125-127, 1982.

A comment to the effect that except for two node models the immediate and the non-immediate blocking policies are not equivalent and it is necessary to insert a blocking indicator into the state description in order to study models with non-immediate blocking policy.

(17) T. Altıok and S. Stidham, Jr. " The allocation of internode buffer capacities in production lines", *IIE Trans.*, **15**, pp. 292-299, 1983.

Model : A tandem line of servers with finite capacity buffers and *independent* and *exponentially* distributed service, up and down times is considered. The *non-immediate* blocking policy is assumed with a *saturated* first node server.

Measure : Optimal allocation of buffer capacities so as to maximize the average output rate is considered.

Method : First, the *effective* service completion time is shown to have a two node Coxian distribution, so that the system can be transformed into a tandem line of *reliable* servers with two node Coxian service distributions. The states of the continuous-time Markov chain are defined and the balance equations are solved by using the power method. A search technique for optimal buffer allocation is also discussed.

(18) S.B. Gershwin and I.C. Schick, " Modeling and analysis of three-node transfer lines with unreliable machines and finite buffers", *Operations. Res.*, **31** , pp. 354-380, 1983.

Model : A tandem system with finite capacity intermediate buffers is considered. The service times are assumed to be *constant* and *equal* while the the up and down times are *geometric* and *independent* from the state of the system. Failures are assumed *operational* and the *non-immediate* blocking policy is adopted with a *saturated* first node server.

Measures : The queue length distribution and the efficiency of the system (production rate) are considered.

Method : The states are defined as in (4) and sum-of-products form solution is assumed for the steady-state probabilities of the internal states, while for the boundary states, expressions are derived by using the transition equations. Therefore the order of the system is reduced from N^2 to N, where N is the total buffer capacity. However, the reduced system is not sparse and may also become ill-conditioned. The results for the general case are applied to a tandem system with three nodes. The internal and the boundary states are identified and a procedure is discussed for the solution of steady-state probabilities.

Results : The following results based on simulations are given for three node systems.

- Average number jobs in the system is inversely proportional (resp. proportional) to the failure rate of the first (resp. third) server.

- The efficiency and the error in the calculations of the steady-state probabilities increases with the total buffer size.

- The efficiency stays approximately constant when all the probabilities are multiplied by a constant number and the buffer capacities are divided by the same number.

- Production rate is not affected by the reversal of the data describing the system, while the error in the numerical calculations is.

- For a balanced line, the maximum efficiency is obtained when the intermediate buffers have equal capacities.

- When the last node server is almost reliable, the system behaves like a two node system.

(19) S.B. Gershwin, "An efficient decomposition method for approximate evaluation of tandem queues with finite storage space and blocking", to appear in *Operations Res.*, 1983, revised 1986.

Model : The model described in (18) is considered.

Measures : The average queue length and the throughput of each node are studied.

Method : An approximate decomposition approach based on the flow conservation is presented and a simple algorithm to calculate the above performance measures is developed. For each buffer the algorithm assumes that the upstream and the downstream parts of the tandem line can be adequately summarized by two servers with geometric up and down times, whence decomposes the system into several two node subsystems. Several numerical examples are provided.

(20) T. Altıok, " Approximate analysis of production lines with general service and repair times and with finite buffers", IE Rep. **84-4**, Rutgers University, 1984.

Model : A tandem line with finite capacity buffers between the servers is considered. The servers are assumed to have *independent Erlang* service time distributions. The up and down time distributions are assumed to be *exponential* and *general*, respectively. The case of a *saturated* first node server and the case where there is a *Poisson* arrival stream to an infinite capacity buffer in front of the first node server are both considered under the *non-immediate* blocking policy.

Measures : Average number of jobs in each buffer and the server utilizations are considered.

Method : Although an expression for the cumulative distribution of service completion time is obtained, it is very complicated to deal with. However, in the case of exponential repair times, by observing the corresponding Laplace-Stieltjes transforms, the service completion time distribution is seen to be the sum of several phase-type distributions with two phases. Then, a cumulative distribution function is obtained by assuming that at most one breakdown may occur during the processing time of a job so that the failures can be incorporated into the service completion times and are approximated by specific phase-type distributions.

This approximation uses the first two or three moments of the derived service time distribution. Particular mixtures of the sum of exponential distributions are chosen by empirical observations. Then, by using the results of Perros and Altıok [15], the effective service time distribution of the i^{th} server is represented by a phase structure involving $2 \times (M - i + 1)$ phases, for $1 \leq i < M$, where M is the total number of nodes in the system. Numerical examples and some empirical observations are also given.

The following two papers discusses approximation algorithms for *assembly/disassembly* networks with blocking. In such networks each buffer is connected to *exactly* one upstream server and one downstream server, while each server is connected to *at least* one buffer. Loops and bypasses are not allowed in both papers. A server that has more than one upstream (resp. downstream) buffer performs assembly (resp. disassembly) operations, and is starved (resp. blocked) if any one of these buffers is empty (resp. full).

(21) S.B. Gershwin, "Modeling and analysis of assembly/disassembly networks", presented at the *IEEE Conference on Systems, Man and Cybernetics,* Atlanta, Georgia, 1986.

Model : An assembly/disassembly system with finite capacity intermediate buffers is considered. The service times are assumed to be *constant* and *equal* while the the up and down times are *geometric* and *independent* from the state of the system. Failures are assumed *operational* and the *non-immediate* blocking policy is adopted with a *saturated* first node server.

Measures : Average number of jobs in each buffer and the server utilizations are considered.

Method : The decomposition algorithm of Gershwin (19) is extended to such networks.

(22) M.B.M. de Koster, "Approximation of assembly/disassembly systems", presented in *TIMS/ORSA Conference*, New Orleans, 1987.

Model : An assembly/disassembly system with finite capacity intermediate buffers is considered when the service times are *constant* but *different* while the the up and down time distributions are *exponential.* It is assumed that the product flow is *continuous* and the first node server is *saturated.*

Measures : The average buffer content and the throughput of each node are considered.

Method : The network is approximated by repeated decomposition and aggregation steps by using the fact that in assembly/disassembly networks the flow direction in an arbitrary branch can be reversed without changing the behavior of the rest of the network. Numerical results and extensions are mentioned.

(23) L. Gün and A.M. Makowski, " Matrix-geometric solution for two node tandem queueing systems with phase-type servers subject to blocking and failures", manuscript, 1987.

Model : A two node tandem system with a finite capacity intermediate buffer is considered in discrete-time when the service and the repair time distributions are both *phase-type*. The *immediate* blocking policy is assumed with a *Bernoulli* arrival process of parameter λ to a finite capacity buffer in front of the first node server.

Measures : The invariant probabilities of the joint queue length distribution are studied.

Method : The effect of failures are first incorporated into the phase structure and the effective service time distribution is shown to be still of phase-type. Necessary and sufficient conditions for the irreducibility of this new distribution are given. Then, the states of the tandem system are identified and the probability transition matrix is given explicitly. The cases $\lambda = 1$ and $\lambda < 1$ are considered separately; the irreducibility of the underlying Markov chains is discussed for each case and the invariant probability vectors are obtained in matrix-geometric form, with the corresponding rate matrices given in terms of the system parameters. Continuous-time formulation is only briefly mentioned and major differences with the discrete-time results are pointed out. Numerical examples are provided.

Author Index